What are comparable corpora?

Belinda Maia
Faculdade de Letras da Universidade do Porto
(FLUP)

1. A definition

The EAGLES - Expert Advisory Group on Language Engineering Standards Guidelines (1996) (see http://www.ilc.pi.cnr.it/EAGLES96/browse.html) gives us the following definition for 'comparable corpora':

> "A comparable corpus is one which selects similar texts in more than one language or variety. There is as yet no agreement on the nature of the similarity, because there are very few examples of comparable corpora".

Although progress has, no doubt, been made in the construction of comparable corpora since 1996, there is very little literature in which the characteristics of comparable corpora are explained and analyzed. It would seem that they are ideals rather than realities. In this paper we shall try to look at the reasons for constructing comparable corpora, the theoretical and practical problems they pose and the uses to which they can be put.

2. The reasons for constructing comparable corpora

The EAGLES report considers that:

> "The possibilities of a comparable corpus are to compare different languages or varieties in similar circumstances of communication, but avoiding the inevitable distortion introduced by the translations of a parallel corpus".

The distinction between parallel and comparable corpora had been made earlier by Baker (1996) and it has become popular, even though some people still prefer to refer to 'aligned' and 'parallel' corpora instead. It is useful in that it draws attention to the fact that translated text necessarily bears a considerable resemblance to the original. Therefore, even a translator who is sensitive to the different conventions applicable to a particular genre in two languages will have difficulty in avoiding Source Text (ST) influence. He/she may find that the problems related to the general lexicon or specific terminology can be solved satisfactorily, and even that the structure of individual sentences can be adapted minimally and yet still convey the message of the ST. However, it is far less easy to change the text and information structure to suit the discourse conventions of the Target Text (TT), if nothing else because translators are not always prepared to cope with the multiplicity of these conventions and because the expectations of the client do not usually include any radical revision of the structure of the text. Even if a publisher or editor favours such interference, the author might well have objections to this, as Munday's (2001:154) reference to attempts by a publisher/ translator team to 'disentangle' the plot in Kundera's *The Joke* demonstrates. Academics writing in languages and for cultures that encourage long sentences and complex argument structure have been known to throw up their hands in horror at attempts by Anglo-American translators to render their texts more 'readable' in the TT.

One can also see how such conventions affect translations by looking at foreign language newspapers and examining the reporting of international news. The conventions of reporting news in the Anglo-American culture encourage a 'factual' style expressed in fairly simple SVO sentence structures, whereas news reporting in cultures which prefer a less direct style, and a degree of hypothesis rather than certainty, will favour a style in which sentences begin with adjuncts or disjuncts, and wide use is made of impersonal expressions. An analysis of reports of international news – which more often than not have been translated hurriedly from material supplied in English by international agencies – will show how the sentence and text

structure obey Anglo-American rather than the local text conventions apparent in the original reporting of news in that language. (See Maia: 1999, 1997 and 1996).

One must also take account of more pragmatic factors influencing the translation process in professional translation. More often than not the client wants something close to the original, believes that translation is a fairly mechanical process, and is easily seduced by the promise of reusable material offered by translation memories. The conditions created by translation memories – the physical alignment of the two texts, repetition of previously translated chunks of language, and uniformity of terminology - all encourage similarity between original and translation.

Translation theorists would do well to remember that most translation is far closer to the original than their theories would suggest. Much of the descriptive research work done on translation in recent years focuses on the odd examples in texts rather than on the aspects of similarity, and literary texts are given more attention than non-literary texts, despite the fact that they constitute approximately only 5% of the translation done per year in the world. However, one could argue for a gradient between parallel and comparable texts in which literary texts are not necessarily the most creative. The 'foreignizing' translations advocated by Venuti's (1995) will probably provide fairly easily alignable parallel corpora, but translators following the recommendations of functionalist theorists like Reiss (1997) and adapting, rather than simply translating, advertisements to a target culture, are arguably providing us with good comparable corpora, if the result is successful.

Comparable corpora are seen as answering perceived needs for texts as examples of 'natural' original text in the source language culture. Although genre analysis is a fairly recent development in the academic world, anyone who lives in a multicultural environment is well aware of the varying textual conventions in different cultures. For example, legal texts are written according to local conventions and often reflect centuries of legal history. The Eurospeak we see in the EC texts is a homogenisation of a variety of legal conventions. Academic and scientific texts also vary considerably according to the different cultural conventions, and there are plenty of socially conventional texts, like 'births, deaths and marriages' announcements, or advertisements for houses and jobs, that have to be adapted rather than translated if they are to function adequately in the target context.

Apart from their theoretical usefulness, comparable corpora have other advantages. One is their availability. It is a lot easier to find original texts on a particular subject than to find a pair consisting of the original + a (good) translation. This means that one can aim for a larger corpus and a greater degree of variety within the corpus, factors that can only lead to greater reliability.

Another advantage of comparable corpora is their versatility. Although parallel corpora have obvious uses for research into the translation process and product, their use for other areas, like terminology extraction, depend very largely on the good quality of the translation and the research done for it by the translator. Comparable corpora, on the other hand, can be applied in a wide range of other research areas, such as Discourse Analysis and Pragmatics. They also offer wider possibilities for Terminology Extraction, Information Retrieval and Knowledge Engineering than parallel corpora.

3. To what extent can texts be 'similar' – and corpora 'comparable'?

To go back to the quotations from EAGLES - "A comparable corpus is one which selects *similar texts* in more than one language or variety" and "The possibilities of a comparable corpus are to compare *different languages or varieties* in *similar circumstances of communication*, but avoiding the inevitable distortion introduced by the translations of a parallel corpus" (my emphasis), one is faced with the inherent – and deliberate – vagueness of the concept of *similarity*. Besides this, one has to pursue this notion of similarity across languages, cultures, text varieties or genres, and circumstances of communication.

First of all, since we are talking about corpora – as opposed to just texts – we must decide on the idea of similarity in relation to both form and content. The form of corpora will be of interest to those who actually construct them. In other words, we shall need to consider the size of the corpus, in terms of number of texts and words, and the nature of the individual texts in terms of words, sentences, and paragraphs. We

shall also need to consider the format. Do we want them in .txt, .doc, .sgml, .html., or .xml, and do we want to eliminate all the formatting, images, etc., or do we consider that these items are necessary for analysing levels of discourse beyond mere text? These are questions that need to be asked, but which will not occupy us here.

The content of corpora is of paramount importance in the construction of comparable corpora, and guidelines on what to look for should be drawn up for this purpose. First of all we should decide whether we wish to aim for general or specialized corpora. Although, of course, the distinction between the two is, in actual practice, fuzzy, the reasons for making it is related to the use we intend to make of the corpus, as we shall see below.

We need to start by looking at the similarity of the texts in relation to both their structure and their function. Is the structure important? For example, are they formal, carefully constructed texts – as with legal texts, or are they informal, loosely organized discourse – as with transcriptions of conversation. Then we need to look at their function in the source culture. Do the texts have a specific social or cultural importance that needs to be taken into consideration?

After this we can look at the texts in terms of *register*, in the Hallidayan sense. We need to decide on the *field* – situation, subject matter, or topic of the texts, and the *tenor* – or the interpersonal relationships involved in the communication situation that lead us to decide on the level formality/informality, politeness, etc. We also need to consider the *mode*. Are the texts examples of spoken language: e.g. speech, formal dialogue, or conversation? Are they written texts: e.g. books, essays, instruction manuals, or novels? Do they involve multimedia, as in the case of hypertext in encyclopaedias like Encarta, films or other media? A further dimension that needs to be considered is that of *dialect*, which can be associated to geographical factors - e.g. urban/rural areas, developed/developing countries; temporal factors - e.g. historical periods and different age groups; or social factors - e.g. social classes and educational backgrounds.

Having established guidelines for similarity of texts, we then need to establish what we understand by comparability in corpora. In theory, this may seem to be fairly obvious. A comparable corpus will consist of a balanced quantity of a certain quality of texts. In practice, it is difficult to arrive at this solution.

If we think of Very Large Corpora, we can only consider them comparable if they are similar in size and constructed according to same criteria, e.g. quantity and quality of text types. Examples of such corpora might be the British National Corpus, the Mannheimer Corpora and other less visible corpora that have served various lexicographical projects. However, apart from the problems of comparable size, one would need to examine the criteria for building these corpora carefully before one could consider them comparable.

Newspaper corpora are popular among corpora makers because of the ease in which they offer a large quantity of texts on a wide variety of subjects. However, newspaper corpora vary considerably according to type - 'quality' or 'popular', content - general or specialised, and time of writing. Now we also have the term 'concurrent' corpora, which can mean corpora collected on the same subject or news item in several newspapers. The date of writing is often an important factor. Well-known newspaper corpora are the CETEMPúblico in Portuguese – see http://www.linguateca.pt and the Reuter's Corpus in English – available for research on CD-ROM.

If we set out to construct a comparable corpus of literary texts, the task becomes extremely complicated, and involves comparisons at both inter-lingual and intra-lingual levels. Just to give an idea of the problems involved, let us think in terms of period - Medieval, 18th Century, or Post-war; School - Romanticism, Realism, or Post-modernism; and Genre - Novel, science fiction, drama, or poetry. And that is not to mention the differences between individual authors. Comparable corpora of this kind have to be tailored to individual research and study needs. Laviosa (1997) describes an interesting attempt at constructing such a literary corpus.

Most of the experiments with comparable corpora so far, however, have been made with specialized texts

of a technical or scientific nature. They take the form of encyclopaedic articles, pamphlets, manuals, textbooks, academic articles and papers, dissertations, and even theses. The chief criteria for their selection are their subject content and their register, and the main objective for constructing them is related to terminology research (see Pearson, 1998 and Bowker & Pearson, 2002).

4. Constructing and using comparable corpora

The first questions to ask when constructing a comparable corpus are: Why do we need a comparable corpus? What shall we do with it? And the next question is - where does one start?

4.1 Comparable corpora – general language

To construct very large comparable corpora from scratch in two or more languages is a mega-proposition and would need an enormous amount of goodwill. However, very large corpora that are organized for academic purposes with the objective of covering a wide variety of language varieties can be considered roughly comparable. The usual function of such corpora is to allow lexicographers and other linguists to establish general patterns of language usage rather than specific genre analysis, so the degree of comparison is very diluted. Very large corpora like the Canadian Hansard and the EC database of texts are parallel corpora.

Smaller general corpora, complete with tagging, like the ICAME corpora (Brown, LOB etc) and similar corpora in other languages, offer the possibility of more controlled comparative and contrastive research into general language at all levels. Newspaper corpora are a popular solution in the quest for 'concurrent corpora'. Examples we have been involved with at FLUP are corpora of war reports, and football during the World Cup, and another possibility would be political texts during election campaigns. One can also compare styles of journalism by comparing individual journalists.

4.2 Comparable corpora – specialized language

The objective of most comparable corpora makers is to concentrate on producing small corpora in specific areas. Making comparable corpora out of texts of fairly general subject matter but similar text type offer various possibilities for Discourse Analysis, Pragmatics, Genre Analysis and Sociolinguistics. Examples of such corpora can include collections of Encyclopaedia entries, tourism pamphlets, literary texts of a similar period, school or genre, and technical and scientific texts with a similar form or function. The possibilities and limitations depend on the objective of the corpus maker. However, the most common comparable corpora in demand are those in special domains. These can be texts at various levels of specialization, but the tendency is to look for texts with a high lexical or terminological density. This is probably due to the fact that, as with general language corpora, the perception of the usefulness of corpora stems from lexicographical, and in this case, terminological interest.

A good deal of work in specialized comparable corpora is of a more informal type. Several people have developed such mini-, do-it-yourself, disposable corpora as teaching aids for specialized translation classes (see Maia 1997 & 2000; Varantola, 2000; Zanettin, 2002). Most have found them very successful and, as training material they have definite advantages. However, the nature of translator training means that we have to combine a good mixture of breadth and depth in subject matter for such corpora. We can use official general corpora as a general language resource, we can encourage students to find texts to solve problems for specific specialised language translations, and we can set individual or group work on collecting small corpora with a view to their a) learning something about the subject b) learning something about the type of text involved and c) extracting useful terminology. The Internet has made this type of methodology possible, and it is become increasingly common practice in classes.

However, one needs a project or more systematic expert support if one is going to go beyond this. One also needs to be able to use texts of a certain quality, and these texts are not always easily available, particularly on the Internet. It is also true that experts do not excited about helping translation studies, but they do react positively if one talks in terms of terminology, databases, information retrieval and other things. It is not usually difficult to explain the potential advantages of proper specialized corpora as a basis for this work,

and my own experience is that domain experts understand the potential of such corpora rather better than my colleagues in Modern Languages.

One must, however, be prepared to specialize seriously. One cannot specialize at the levels of Geography, Engineering or Medicine, nor even at the levels of population geography, mechanical engineering or oncology. One must go at least one level deeper and work on terminology for discussing ethnic minorities (at FLUP), tribology (at FLUP) or breast cancer (see Faber, 2002). For this one needs – ideally – corpora of introductory books to the subject aimed at university students who are already studying in one of the departments of the Faculties of Geography, Engineering or Medicine. This type of book usually contains most of the terminology they will need to discuss the special domain, together with pedagogically phrased definitions. After this, one should collect international standards documents, academic articles, dissertations and theses.

4.3 Constructing comparable corpora

All corpora construction must establish an overall general policy in relation to both the form or computational structure, and the content of the sub-corpora. There should also be a policy about the availability of the corpora to a general or restricted public. Although anything available in digital form is a candidate for a corpus, it is more economical to build specialized sub-corpora for specific objectives. Corpora construction must take into account a variety of factors, of which copyright is probably the most difficult to deal with. However, when goodwill and sensible computational protection come together so that texts can be used without endangering the author's rights, a lot can be achieved.

Choosing texts for comparable corpora is not easy and the process requires both good luck and good judgement. One also needs to decide where on the scale from lexicography/terminology to discourse analysis one wants to carry out research. At the terminology end, one needs texts by a recognized expert and rich in terms. At the discourse level one needs to take into account the effect of external factors on the text, such as the conventions of writing such texts in specific cultural or social situations and the idiosyncrasies of individual authors.

With all comparable corpora of the kind being discussed, one must be aware of the homogenising effect of internationalisation, Eurospeak and the Anglicisation of scientific terminology. As one who has had the experience of trying to collect comparable corpora at a certain level in English and a less international language – European Portuguese, I have had to face the fact that it is difficult to find suitable texts in the latter. International publishing has made it far easier and cheaper for a publisher to translate and adapt a best selling Anglo-American textbook or encyclopaedia than to provide the local talent with the opportunity and financial backing to show what they are worth. Most international documents now start their lives in English, or at least in some form of international English agreed upon by all. If one adds to this the fact that many people for whom English is a second language publish in English, and even follow American guidelines for writing texts, one can understand why it is so difficult to find texts that have not been contaminated by Anglo-American terminology and text conventions.

4.4 Annotation of comparable corpora

Comparable corpora can be analysed using a variety of techniques, including various types of annotation, keywords and multi-word extraction tools. Research at the terminology end of the spectrum tends to revolve around devising techniques that extract terms both rapidly and efficiently in terms of both precision and recall. In order to be useful for syntactic or discourse analysis, corpora need proper tagging and analysis before work can begin – and not everyone has this kind of facility easily available. Computer-manageable theories need to be developed for different languages to cope with such research. Every language needs a different POS tagger, for example, and annotation beyond part-of-speech tagging will be subject to arguments over a wide variety of theoretical criteria, and limited by the practical possibilities of applying these criteria in a way that allows rapid annotation of large quantities of text (see: Garside et al, 1997, and van Halteren, 1999).

To give an example of the problems of annotation, a group of us at FLUP are at present working on an experiment with our Master's students on evaluating machine translation from English to Portuguese. We use the British National Corpus (BNC) and Google to find interesting natural examples in English to translate, and then submit the examples to an on-line tool we have developed that allows us to request translations from several of the free on-line MT programmes simultaneously. The students, who are translators, rather than experts in linguistics, describe the phrase or structure they are trying to test using the POS annotation of the BNC, and then analyse the results in terms of a simplified version of the IMS corpus workbench annotation, as adapted for Portuguese by Diana Santos for the AC/DC corpora of the Linguateca project. The objective of using different annotation schemes is to allow for both original and translation to be discussed in terms suited to the individual languages, rather than aim for some universal set of categories. Discussion of this type of exercise is, I presume, is one of the objectives of this workshop, and we hope to gain from the experience of others.

5 An ongoing project

The overall objective of the Linguateca project is the computational processing of Portuguese. Although researchers in the four universities involved are working on a variety of topics, the more immediate objectives of the Universidade do Porto, and those which are of interest to this workshop, are:

- to construct the necessary computational tools for using a variety of corpora for research and pedagogical purposes;
- to construct comparable corpora in Portuguese and English;
- to contribute to the already existing COMPARA corpus of parallel texts;

Although we expect to help colleagues, researchers and students to use and create corpora for their own use, charity begins at home, and the immediate focus of our energies is to support the students of our Master's degree in Terminology and Translation (see: http://www.letras.up.pt/translat/i_mepr.htm for the programme). We have ongoing dissertations that require our support to organize their corpora, provide technical assistance for semi-automatic term extraction and help create multi-dimensional conceptual frameworks and ontologies. For this we already have texts in special domains and intend to add further corpora – both comparable and parallel - as and when the opportunity arises.

Our policy is to start from specific situations and work towards filling in a more general framework. The emphasis is on finding good quality texts that allow us to develop reliable terminology databases, as well as allowing experimental work in information retrieval and discourse analysis. In all cases we have the cooperation of domain experts. Experience tells us that we shall probably have to compromise with our ideal of making these corpora widely available, but our objective is to obtain copyright permission for as big a group of users as possible.

To give an idea of what is already happening, here are a few of our on-going projects. We have one group of students working with a colleague from the Geography department to produce a dictionary + database of terminology in the area of Population Geography, with related dissertations on the creation of a multi-dimensional conceptual framework and ontology of the area, and on the multimedia texts resulting from sub-titling and dubbing in documentaries related to this area. One student is preparing a dissertation using semi-automatic term extraction in the domain of marketing, and another is using concurrent corpora from newspapers during the World Cup to study metaphors in football.

Future work with the Geography department is planned in the domain of 'natural disasters', possibly with the cooperation of insurance specialists. We also plan to prepare a Portuguese version of Delisle, Lee-Jahnke and Cormier's *Terminology of Translation*, with the help of our colleagues in Translation, as well as Bert Esselink's *Introduction to Localization*, under the guidance of the computer engineers in the Faculty of Engineering.

Our longer-term objectives are to extend the notion of comparability to genre-specific corpora and restricted general language corpora, and to construct integrated networks of comparable corpora and to extend these objectives to other languages as well. Quite obviously, it would be interesting to also coordinate our work with similar projects elsewhere.

6 Conclusions

How far, therefore, can any two or more corpora be comparable? It should be clear from what has been said above that, to a certain degree, comparability is in the eye of the beholder. Each corpus project results from perceived needs, and only succeeds if one combines good luck in finding appropriate texts, good judgement in selection and deciding how to best use them, and good computational tools for analysis. Building and using comparable corpora offer a theoretical challenge that is truly fascinating and, although the practicalities of corpora making may cause us to fall short of our ideal, we should not lose sight of the wider implications. The work of the FLUP group of the Linguateca project is at present focused primarily on areas that will benefit our ongoing research into terminology and translation. However, some of this work already has implications involving information retrieval and other things of interest to knowledge engineering. Inter-disciplinary cooperation is a major characteristic of our work and everyone involved seem to find this intellectually stimulating.

Others, too, are beginning to see the possibilities of other kinds of comparable corpora. My colleague from psycholinguistics is making a corpus of 'chat', and, no doubt, this could be compared with similar corpora of e-mail messages and chat. The theoretical implications are quite different from those that have guided the other work discussed here, but there are definitely prospects for interesting research there.

Bibliography

Baker, M 1996 Corpus-based translation studies - the challenges that lie ahead. In Somers, H.L (ed.), *Terminology, LSP and Translation*. Amsterdam/Philadelphia, John Benjamins, pp 175-186.

Bernardini, S, Zanettin, F (eds) 2000 *I corpora nella didattica della traduzion.* Bologna, CLUEB.

Bourigault, D, Jacquemin, C, L'Homme M-C (eds.) 2001 *Recent Advances in Computational Terminology*. Amsterdam/Philadelphia, John Benjamins Publishing Co.

Bowker, L, Pearson, J 2002 *Working with Specialized Language- a practical guide to using corpora.* London/ New York, Routledge.

Charlet, J, Zacklad, M, Kassel, G, Bourigault, D 2001 *Ingénierie des connaissances.* Paris, Éditions Eyrolles.

Delisle, J, Lee-Jahnke, H, Cormier, M.C 1999 *Terminology of Translation.* Amsterdam/Philadelphia, John Benjamins Publishing Co.

Esselink, B 2000 *Introduction to Localization*. Amsterdam & Philadelphia, John Benjamins Publishing Co.

Faber, P 2002 Terminographic definition and concept representation, in Maia, B, Haller, J, Ulrych, M (eds.) pp. 343-354.

Garside, R., Leech, G, McEnery, A 1997 *Corpus Annotation – Linguistic Information from Computer Text Corpora,* London/New York, Longman.

Halliday, M.A.K. 1984 *An Introduction to Functional Grammar*. London: Edward Arnold.

van Halteren, H (ed.) 1999 *Syntactic Wordclass Tagging.* Dordrecht/Boston/London, Kluwer Academic Publishers.

Klaudy, K & Kohn, J (eds.) *Transferre Necesse Est - Proceedings of the 2nd International Conference on Current Trends in Studies of Translation and Interpreting 5-7 September, 1996, Budapest, Hungary*. Budapest, Scholastica.

Laviosa, S 1997 How Comparable Can 'Comparable Corpora' Be? In *Target* 9(2), pp 289-319.

Lewandowska-Tomaszczyk, B, Melia, P.J (eds.) *PALC '97 Practical Applications in Language Corpora,* Lodz, Lodz University Press.

Maia, B, Haller, J, Ulrych, M (eds.) 2002 *Training the Language Services Provider for the New Millennium.* Porto, Universidade do Porto.

Maia, B 2000 Making corpora: a learning process. In Bernardini, S. & F. Zanettin, (eds.) pp 47-6.

Maia, B 1999 'Natural' sentence structure in English and Portuguese and its influence on the organisation of information in the process of translation. In Pinto, M. de G, Veloso, J, Maia, B (eds.), pp 603-6.

Maia, B 1997 Do-it-yourself corpora ... with a little bit of help from your friends. In Lewandowska-Tomaszczyk, B, Melia, P.J (eds.) pp. 403-410.

Maia, B 1997 Sentence Structure and Thematization in Comparable and Parallel Texts. In Klaudy, K, Kohn, J (eds) pp 541-547.
Maia, B 1996 The sentence as a unit of translation. In the *Proceedings of the III Jornadas de Tradução do ISAI*. Porto: ISAI, pp 27-36.
Munday, J. 2001 *Introducing Translation Studies: Theories and Applications.* London/New York, Routledge.
Pearson, J. 1998 *Terms in Context*. Amsterdam/Philadelphia, John Benjamins Publishing Company.
Pinto, M. de G, Veloso, J, Maia, B 1999 *Proceedings of ISAPL '97 – 5th International Congress of the International Society of Applied Psycholinguistics.*
Somers, H.L (ed.) 1996 *Terminology, LSP and Translation*. Amsterdam/Philadelphia, John Benjamins Publishing Company.
Varantola, Krista. 2000 Translators, Dictionaries and Text Corpora. In Bernardini, S. & Zanettin, F (Eds). 2000, pp 117-133.
Veronis, J (ed) 2000 *Parallel Text Processing: Alignment and Use of Translation Corpora.* Dordrecht, Kluwer Academic Publishers.
Zanettin, F 2002 DIY corpora: the WWW and the translator. In Maia, B, Haller, J, Ulrych, M (eds.), pp 239-248.

Internet links – last accessed 5 March 2003

British National Corpus: http://www.hcu.ox.ac.uk/BNC/
EAGLES - Expert Advisory Group on Language Engineering Standards Guidelines (1996): http://www.ilc.pi.cnr.it/EAGLES96/browse.html
Faculdade de Letras, Universidade do Porto – Master's in Terminology and Translation: http://www.letras.up.pt/translat/i_mepr.htm
Google: http://www.google.com
IMS Corpus Workbench: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/
Linguateca: http://www.linguateca.pt